Reffay, C., Teplovs, C., Blondel, F.-M. (2011). Productive re-use of CSCL data and analytic tools to provide a new perspective on group cohesion. In *Connecting Computer-Supported Collaborative Learning to Policy and Practice: CSCL2011 Conference Proceedings. Volume II – short papers*, Hans Spada, Gerry Stahl, Naomi Miyake, Nancy Law (Eds.), pp. 846-850, 4-8 july, 2011, Hong Kong, China.

Productive re-use of CSCL data and analytic tools to provide a new perspective on group cohesion

Christophe Reffay, STEF, INRP - ENS Cachan - UniverSud, Cachan, France. christophe.reffay@inrp.fr
Christopher Teplovs, University of Toronto, Ontario, Canada. chris.teplovs@utoronto.ca
François-Marie Blondel, STEF, INRP - ENS Cachan - UniverSud, Cachan. francois-marie.blondel@inrp.fr

Abstract: The goals of this paper are twofold: (1) to demonstrate how previously published data can be re-analyzed to gain a new perspective on CSCL dynamics and (2) to propose a new measure of social cohesion that was developed through improvements to existing analytic tools. In this study, we downloaded the Simuligne corpus from the publicly available Mulce repository. We improved the Knowledge Space Visualizer (KSV) to deepen the notion of cohesion by using a dynamic representation of sociograms. The Calico tools have been used and extended to complete this cohesion measure by analyzing lexical markers. These complementary analyses of cohesion, based on clique sizes and communication intensity on the one hand, and lexical markers on the other hand, offer more detailed information on (a) the relationships between participants and (b) the structure and intensity of communication. In particular, the analyses highlight strong convergences that were not visible in the previous analysis.

Introduction

Because of their complexity, authentic learning experiences are hard to replicate. This makes comparison and validation of research tools, methods and results in CSCL difficult. Research collaboration has been well advocated in the context of Technology Enhanced Learning in order to make a greater impact and further elevate our research quality (Chan et al., 2006). This issue has been addressed by various projects that have been concerned with data sharing within communities of researchers.

In the research data sharing perspective, the Dataverse Network project (http://thedata.org/) described by King (2007), shows why datasets have to be shared, or at least identified and recorded as persistent, authorized, and verifiable data. For the Intelligent Tutoring Systems (ITS) field, the PSLC DataShop (Koedinger et al., 2010) provides a data repository including data sets and a set of associated visualization and analysis tools in order to evaluate the action/feedback interaction between learners and (virtual) tutor tools. In the CSCL community, the DELFOS framework (Osuna, Dimitriadis, & Martínez, 2001) provides an XML based data structure (Martínez, de la Fuente, & Dimitriadis, 2003) for collaborative actions in order to promote interoperability (between analysis tools), readability (either for human analysts and automated tools) and adaptability to different analyzing perspectives. Some of these authors joined the European research project reported in (Martínez, Harrer, & Barros, 2005) and provide a technical template describing IA tools and a common format.

The Mulce project (http://mulce.org) developed a platform (http://mulce.univ-bpclermont.fr:8080/PlateFormeMulce/) (Reffay & Betbeder, 2009) to share learning and teaching corpora. This new possibility should deepen our understanding of well-contextualized situations and hopefully better validate tools and have a greater impact on the real world of (collaborative online) learning. Even if more than 30 complex objects are already publicly available on this repository, there is still no evidence of productive re-use of these corpora.

The purposes of this paper are (1) to demonstrate how previously published data can be re-analyzed to gain a new perspective on CSCL dynamics and (2) to propose a new measure of social cohesion that was developed through improvements to existing analytic tools.

Social Network Analysis in CSCL

Social interactions are an inherent aspect of CSCL. Considering participants as a social network (Wellman, 2001) provides a framework that can help us understand what are often complex patterns of interaction. Several studies have used techniques from social network analysis to examine patterns of interaction among CSCL participants (de Laat, Lally, Lipponen, & Simons, 2007; Liao, Li, Wang, Huang, & Zhang, 2007; Martínez, Dimitriadis, Rubia, Gomez, & de la Fuente, 2003; Nurmela, Lehtinen, & Palonen, 1999). They suggest that social network analysis (SNA) can provide useful tools in situations where traditional, statistical methods may not be suitable or may obscure interesting results. Wang and Li (2006) provide a brief history of social network analysis and its application to CSCL.

Among the variety of well established measures like indegree, outdegree, centrality, betweenness, density and cohesion, this paper focuses on the latter. Our cohesion measure is based on the analysis of cliques (i.e. subset of individuals in which all persons are connected to each other), k-cliques (i.e. a clique of k

members) and cliques of level n (i.e. in valued graphs: subset in which all individuals are connected to each other, with an edge which value is at least n).

We were interested in re-examining a data set that had been previously used for a social network analysis. Reffay & Chanier (2003) analyzed the data set described in the next section in terms of cohesion. After providing a description of the data we describe how two existing tools were modified to facilitate the development of a more sophisticated measure of cohesion. This analysis, based on cliques, is also compared with a different method using Calico tools and lexical markers.

The Simuligne data re-used

Simuligne is a distance French as a foreign-language learning situation in a trans-disciplinary research project. The global simulation method was generally used for intensive face-to-face language learning courses. In the Simuligne learning situation, it has been adapted to this extensive online learning situation in parallel in 4 basic groups. Everybody worked at a distance; none of the learners had ever met before Simuligne. The participants consisted of 40 learners (English adults in professional training, registered at the Open University, UK), 10 natives (French teacher trainees from the Université de Franche-Comté, Besançon, FR), 4 tutors (teachers of French from the Open University) and one (French) pedagogical coordinator. All agents were dispatched into four basic learning groups, namely: *Aquitania*, *Lugdunensis*, *Narbonensis* and *Gallia*. Each of these groups consisted of 10 learners, two or three natives and one dedicated tutor.

Three groups out of four achieved the simulation, which is a high ratio in distance learning. On May 31st, the *Lugdunensis* group broke up and its two most active learners were transferred to *Aquitania* group. In this study, we focus on the forum exchanges in the four basic groups only for the period before the *Lugdunensis* group broke up, i.e. from April the 3rd to May the 31st.

The Knowledge Space Visualizer

The Knowledge Space Visualizer (KSV) is a software tool (http://chris.ikit.org/ksv/) that facilitates the exploration of social and semantic networks in data collected from online discourse environments (Fujita & Teplovs, 2010). In the current study the KSV was modified to allow the representation of social links between authors from the Simuligne data set based on the number of posts that each pair of authors had read (opened) of each other.

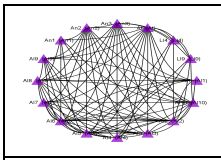
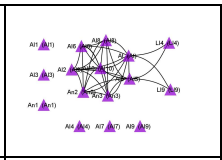
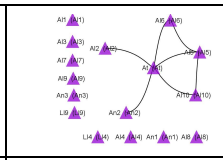
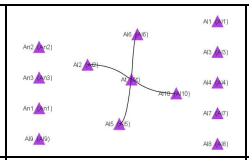
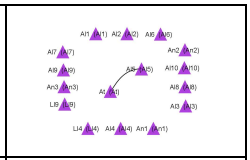
Calico tools to analyse computer mediated discussions

The Calico platform (http://www.crashdump.net/calico/) was developed for sharing and analyzing discussion forum objects (Giguet *et al.*, 2009). The Calico workspace provides several ways to display the contents of messages, to compute quantitative and qualitative indicators about authors, interactions and topics and to display global or local views on messages and topics (http://www.stef.ens-cachan.fr/calico/en/tools.htm). For the purpose of our analysis on specific lexical markers, two Calico tools, namely Colagora and Bobinette, were used to give both general and local measures and views on the utterances of these markers. Colagora displays word occurrences and highlights every matching word in the messages with colors linked to the topics defined by the user. Bobinette is a viewer designed to facilitate reading large forums. It displays messages as circles on a grid with threads in lines and days in columns. Bobinette computes statistics about word topics for each post, thread and day, and highlights messages with the same coding scheme as Colagora.

Social Network Analysis using an adapted version of KSV

The KSV gave us the opportunity to observe the formation of cliques across all the possible interaction intensity values. Considering the graph where nodes represent actors (learners, tutors, natives) and edges communication between actors, KSV draws edges which values are greater than a given intensity threshold and reshapes the graph layout automatically. In Reffay & Chanier (2003), this threshold was fixed and cliques of each group were built with UCINET and compared for this value. The KSV allows us to explore all intensity values for each group and try to find some patterns. This exploration (illustrated on *Aquitania* in table 1) led us to consider the core/periphery model of Borgatti & Everett (1999) and more specifically the Freeman star (Freeman, 1979).

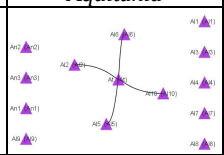
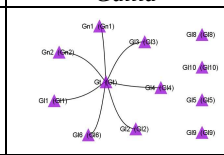
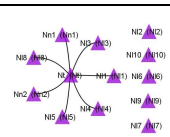
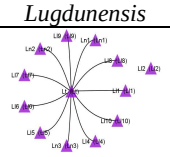
Table 1: Sociograms visualization of Aquitania across intensities with KSV for the first 8 weeks of Simuligne.

				
Intensity=1	Intensity=10	Intensity=100	Intensity=129	Intensity=200

We were interested in examining the following questions in relation to the appearance of the star structure in response to varying the threshold: What is the intensity value? Who is at the center of the star? Who is a branch of it? Who is not connected?

Table 2 shows the following characteristics: (1) each group shows a single well formed star (with more than 2 branches); (2) among the 4 basic groups, *Gallia* and *Narbonensis* are very similar and *Aquitania* and *Lugdunensis* very different for all values; (3) the center of each group's star is the tutor; and (4) *Aquitania's* star is the only one where no native appear in the star branches. The threshold value is a good indicator of the intensity of the exchanges between members for each group. Extreme values are 12 and 129 respectively from *Lugdunensis* and *Aquitania*.

Table 2: Star shapes and thresholds for 4 basic groups for the first 8 weeks of the Simuligne learning session.

	Aquitania	*Gallia*	*Narbonensis*	*Lugdunensis*
Star shape				
Intensity	129	36	49	12
Nb of branches	4	7	7	11
Who is the center?	Tutor	Tutor	Tutor	Tutor
Who is around?	4 Learners	5 Learners, 2 Natives	5 Learners, 2 Natives	8 Learners, 3 Natives

In Figure 1, we show 4 curves (one for each group). The vertical axis represents the intensity and the horizontal one the maximum size (k) of cliques. The first point of *Aquitania's* curve is (k=3, intensity=128). This means that the highest intensity reached by any 3-clique in *Aquitania* is 128. That is, the communication intensity for any subgroup of 3 members is bounded by 128 messages exchanged by pairs. We can see that the value of the star's threshold (from Table 2) corresponds to the top of the curve for each group on Fig. 1.

In order to get a more precise view of these curves for k-cliques with k≥5, the scale of intensity has been changed from Figure 1a to Figure 1b. We can observe that the intensity of the *Aquitania's* internal kernel (up to 4-cliques) is twice greater than the second one (*Narbonensis*). Up to 7-cliques, *Aquitania* shows the highest intensity. But for bigger cliques (k>7), *Gallia's* intensities dominate the graph. It may be that the very high intensity of exchanges between the core members in *Aquitania* and *Narbonensis* was discouraging the peripheral members. The more modest amount of communication between core members of *Gallia* seems to have kept more members in the core.

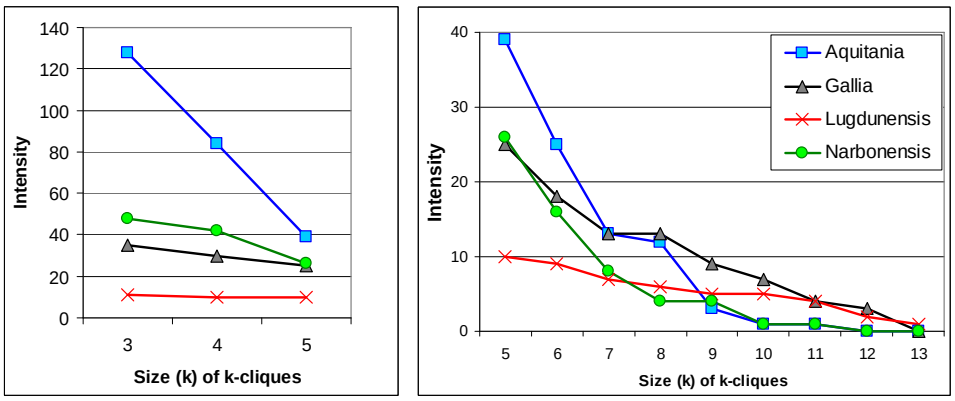


Figure 1a & 1b: Maximum intensity of exchange for each size of cliques on the 4 groups.

Finally, this analysis of cliques for the 4 basic groups shows that (1) *Lugdunensis* cliques (even small ones) have very low intensity values, (2) *Aquitania* and *Narbonensis* have very similar clique characteristics across intensity: a restricted core very active and very little communication exchanged in medium and large cliques. *Gallia's* core (small cliques) shows a lower intensity but this is the group where medium and large cliques have the more intensive exchanges.

The next part analyses the cohesion on the same data, by using lexical markers (provided by Calico). The discussion will show convergences and discrepancies between SNA and lexical analysis of cohesion.

Use of Calico to analyze cohesion through pronoun markers

In their review of text analysis approaches in the social sciences, Pennebaker, Mehl, & Niederhoffer (2003) discuss the links between several linguistic markers like prepositions, pronouns, emoticons, affective words, and social interaction. Yates (1996) suggests that participants in on-line discussions use first and second pronouns more often than in usual written communication. When examining interactivity in discussion groups, Rafaeli & Sudweeks (1997) found that about 25% of the messages they qualified as "interactive" contain first-person plural pronouns, which is significantly greater than the percentage calculated for the entire corpora of messages (about 10%). Following the same technique, we assumed that pronouns markers may be used as an indicator of group cohesion. For the purpose of this study, we counted the "first-person singular" (FPS), "second-person plural" (SPP) and "first-person plural" (FPP) markers in the 4 basic groups. It should be noticed that in French the second-person plural ("vous") is different from the second-person singular ("tu"). Figures from the *Lugdunensis* group should be considered with caution because of the low number of messages.

The frequency of the "first person singular" (FPS) markers is very high for all groups, as already observed by Yates (1996) ; from 70% to 81% of messages contains at least one FPS. Except for the *Lugdunensis* group, the three groups have FPP values that are very similar to those found by Rafaeli and Sudweeks. We can assume that the participants of these groups are in a similar situation because they are invited to interact with each other in the same group. The lower percentage of FPP values in *Lugdunensis* can be interpreted as a possible indicator of lower group cohesion (note that the number of messages of this group is also the lowest).

Table 3: Messages and lexical markers in the four basic groups.

	messages	% of messages with FPS (I)	% of messages with SPP (you)	% of messages with FPP (we)
Aquitania	348	77%	30%	24%
Gallia	159	81%	51%	20%
Narbonensis	175	77%	29%	25%
Lugdunensis	73	70%	41%	18%

We also noticed that the use of "second-person plural" markers (SPP) is different among groups and higher in the *Gallia* group. Using Bobinette we explored and visualized which actors and what messages contain the most significant number of markers in different groups. Looking more closely at the number of messages for each actor, we found that the 3 well-functioning groups contained at least 50% of messages posted by learners, in comparison with the only 36% of messages posted by learners in the *Lugdunensis* group. The part of messages posted by natives is more important in *Gallia* (16%) than in *Aquitania* (7%) and *Narbonensis* (5%).

By facilitating the selection of authors, Bobinette shows that tutors from *Aquitania* and *Narbonensis* wrote similarly high numbers of FPP (80) in comparison to the only 23 FPP written by the *Gallia's* tutor. Furthermore, this abundance of FPP for both of these groups is concentrated in the tutors' messages, 58% for *Aquitania*, 72% for *Narbonensis*, but only 44% for *Gallia*. These results suggest a strong similarity between *Aquitania* and *Narbonensis*, which was not visible in the Reffay & Chanier (2003) analysis.

Discussion

Overall, an important improvement has been made to the earlier cohesion analysis on the Simuligne experiment. Reffay & Chanier (2003) selected a given intensity and drew the cliques only for that value. They suggested the use of hierarchical cluster analysis to find the appropriate intensity value. KSV allows us to examine the entire range of intensity values for clique analysis and it is no longer necessary to choose a fixed intensity value. Instead, we can look for a particular pattern (e.g. a star) and determine the corresponding intensity value.

Two different techniques (SNA and lexical markers) have been used to characterize group cohesion in the same data. These analyses corroborate each other. They both conclude that intensity of exchanges and number of messages are very low for the *Lugdunensis* group to be considered. Lexical analysis shows that the use of "we" is similar in *Aquitania* and *Narbonensis* groups, and both groups also show similar cliques structures across intensity. Besides pronouns, other lexical markers like emoticons, prepositions, and conjunctions could be used with Calico tools to analyze group cohesion.

The Mulce platform facilitated the reuse of the Simuligne corpus to show that comparison of methods and tools on existing data and analysis is possible. Admittedly the process of data reuse and tool modification was somewhat easier than can typically be expected because the data provider worked with the tool developer. Other researchers have found that Simuligne data and context were described with sufficient details to be able to understand them. We demonstrated in this paper that this reuse is productive by bringing more sophisticated indicators and substantial improvements to existing analysis tools (Bobinette and KSV). In this sense, this work opens new perspectives on data reuse in CSCL.

References

Borgatti, S. P., & Everett, M. G. (1999). Models of core/periphery structures. *Social Networks, 21*, 375-395.

Chan, T., Roschelle, J., Hsi, S., Kinshuk, Sharples, M., Brown, T., et al. (2006). One-to-one technology-enhanced learning: An opportunity for global research collaboration. *Research and Practice in Technology Enhanced Learning, 1*(1), 3-29.

de Laat, M., Lally, V., Lipponen, L., & Simons, R.-J. (2007). Investigating patterns of interaction in networked learning and computer-supported collaborative learning: A role for Social Network Analysis. *International Journal of Computer-Supported Collaborative Learning, 2*(1), 87-103.

Freeman, L. C. (1979). Centrality in social networks conceptual clarification. *Social Networks, 1*(3), 215-239.

Fujita, N., & Teplovs, C. (2010). Software-based scaffolding: Supporting the development of knowledge building discourse in online courses. In K. Gomez, L. Lyons & J. Radinsky (Eds.), *Learning in the Disciplines: Proceedings of the 9th International Conference of the Learning Sciences (ICLS 2010) – Volume 1, Full Papers* (pp. 1056-1062). Chicago, IL: International Society of the Learning Sciences.

Giguet, E., Lucas, N., Blondel, F.-M., & Bruillard, É. (2009). Share and explore discussion forum objects on the Calico website. In A. Dimitracopoulou, C. O'Malley, D. Suthers & P. Reimann (Eds.), *Computer Supported Collaborative Learning Practices: CSCL 09 Community Events Proceedings* (pp. 174-176): International Society of the Learning Sciences.

King, G. (2007). An Introduction to the Dataverse Network as an Infrastructure for Data Sharing. *Sociological Methods & Research, 36*(2), 173-199.

Koedinger, K., Baker, R. S. J. d., Cunningham, K., Skogsholm, A., Leber, B., & Stamper, J. (2010). A Data Repository for the EDM community: The PSLC DataShop. In C. Romero, S. Ventura, M. Pechenizkiy & R. S. J. d. Baker (Eds.), *Handbook of Educational Data Mining.* Boca Raton, FL: CRC Press.

Liao, J., Li, Y., Wang, J., Huang, R., & Zhang, Z. (2007). A systemic model of interaction analysis in CSCL. In V. Uskov (Ed.), *Proceedings of the 10th IASTED International Conference on Computers and Advanced Technology in Education* (pp. 463-469). Anaheim, CA: ACTA Press.

Martínez, A., de la Fuente, P., & Dimitriadis, Y. (2003). Towards an xml-based representation of collaborative action. In B. Wasson, S. Ludvigsen & U. Hoppe (Eds.), *Designing for change in networked learning environments: Proceedings CSCL'2003. (pp. 379-383).* Bergen, Norway: Kluwer Academic Publishers.

Martínez, A., Dimitriadis, Y., Rubia, B., Gomez, E., & de la Fuente, P. (2003). Combining qualitative evaluation and social network analysis for the study of classroom social interactions. *Computers & Education, 41*(4), 353-368.

Martínez, A., Harrer, A., & Barros, B. (2005). *Library of Interaction Analysis Tools. Deliverable D.31.2 of the JEIRP IA (Jointly Executed Integrated Research Project on Interaction Analysis Supporting Teachers & Students' Self-regulation).* .

Nurmela, K., Lehtinen, E., & Palonen, T. (1999). Evaluating CSCL log files by social network analysis. . In C. Hoadley & J. Roschelle (Eds.), *CSCL 1999*. Palo Alto, CA: Stanford University.

Osuna, C., Dimitriadis, Y., & Martínez, A. (2001). Using a Theoretical Framework for the Evaluation of Sequentiability, Reusability and Complexity of Development in CSCL Applications. In P. Dillenbourg, A. Eurelings & K. Hakkarainen (Eds.), *Proceedings of EuroCSCL 2001*. Maastricht, NL.

Pennebaker, J. W., Mehl, M. R., & Niederhoffer, K. G. (2003). Psychological Aspects of Natural Language Use: Our Words, Our Selves. *Annual Review of Psychology, 54*(1), 547-577.

Rafaeli, S., & Sudweeks, F. (1996). Networked Interactivity. *Journal of Computer-Mediated Communication, 2*(4).

Reffay, C., & Betbeder, M.-L. (2009). *Sharing Corpora and Tools to Improve Interaction Analysis*. Proceedings of the 4th European Conference on Technology Enhanced Learning: Learning in the Synergy of Multiple Disciplines. (pp. 196-210).

Reffay, C., & Chanier, T. (2003). How social network analysis can help to measure cohesion in collaborative distance-learning. In B. Wason, S. Ludvigson & U. Hoppe (Eds.), *Designing for change in networked learning. Proceedings of the international conference on Computer Supported Collaborative Learning 2003.* (pp. 343-352). Bergen, Norway: Kluwer Academic Publishers.

Wang, Y., & Li, K. (2006). An application of social network analysis in evaluation of CSCL. In R. Mizoguchi, P. Dillenbourg & Z. Zhu (Eds.), *Learning by effective utilization of technologies: Facilitating intercultural understanding*: IOS Press.

Wellman, B. (2001). Computer networks as social networks. *Science, 293*, 2031-2034.

Yates, S. J. (1996). Oral and Written Linguistic Aspects of Computer Conferencing: A Corpus Based Study. In S. C. Herring (Ed.), *Computer-Mediated Communication. Linguistic, Social and Cross-Cultural Perspectives* (pp. 29-46). Amsterdam, NL: John Benjamins.